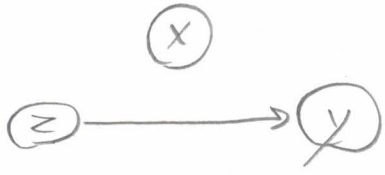
MATH 303 LEC 19

Now let's put the x into the DAG. For the setting of prediction (282) and usually in inference (391) eg. in experimental studies, the x's must be measured before the response is measured. Let's look at some very simple cases

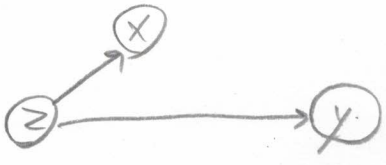
A



Here x does not cause y and x is not correlated to y.

However, a spurious correlation may occur. A model $g(x)$ would perform worse than g_0 since it is overfit (est. err).

B



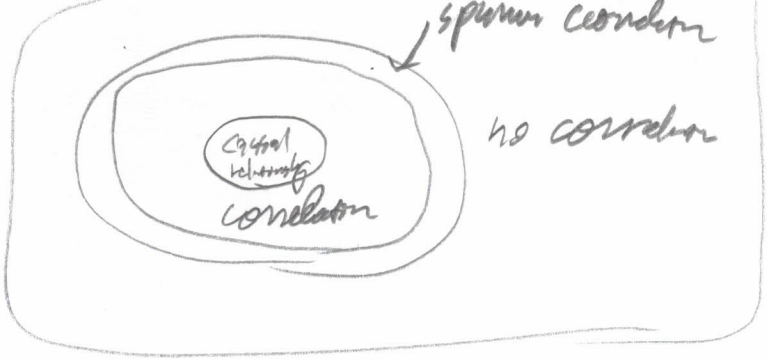
Here x does not cause y but x is correlated to y.

Here, $g(x)$ should do better than g_0 depending on how strong the relationship $X = k_{x,z}(z) +$ ignorance is and how strong $y = k_{y,z}(z) +$ ig carefully it is captured when the model $y \sim x$ is assumed

C

X both causes y and x is corr. to y. $g(x)$ should do better than g_0 depending on how strong the relationship is in $Z = k_{z,x}(x) +$ ig and $y = k_{y,z}(z) +$ ig and how carefully it is captured when model $y \sim x$ is

all possibilities



Thus the full picture, for various $X, Y \ldots$ the relationship between

Thus correlation $\not\Rightarrow$ causation, but sometimes they coincide.

How to prove causation? Thus the big topic to be explored.

Thus causation $\Rightarrow$ correlation.

correlation $\Rightarrow$ causation sometimes i.e. through a common effect (which is called a confounder) or "lurking variable"

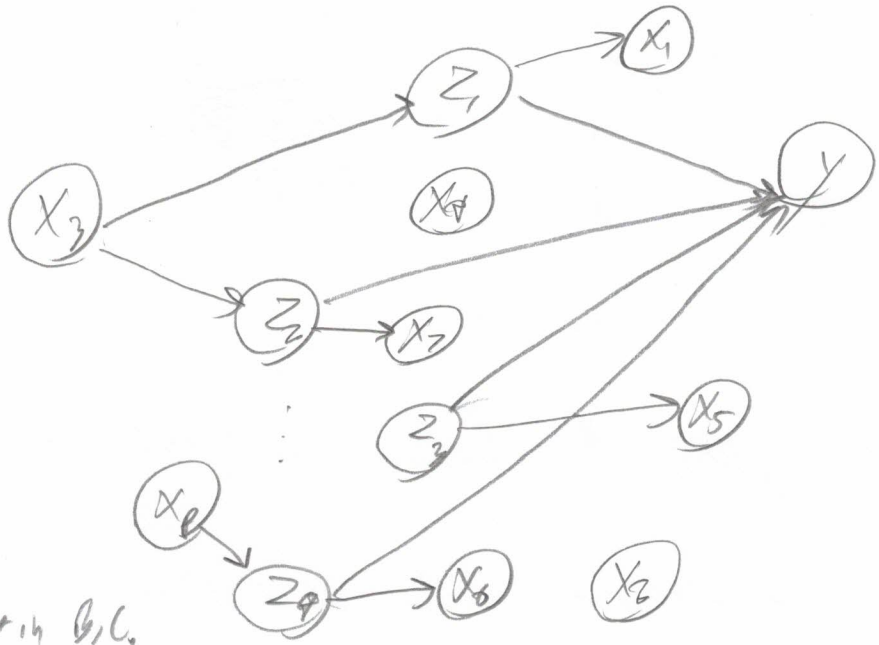
In [13], z is the "confounder" of the relationship of x on y.

If I have more time, the definition will be more exact.

What does the real picture look like with multiple x's?

$y = f(z_1, \ldots, z_q)$

$y = f(x_1, \ldots, x_p) + \varepsilon$



Each scenario A, B, C are represented and more!

You don't want process in A (these useless and will cause overfitting). But you want in B, C.

Given this very messy picture, what does $\hat{\vec{b}}$ mean in the context of OLS, logistic regression, poisson regression, weibull, cox ph, etc?

let's be careful and build up an interpretation iteratively

OLS / robust regression $\quad \hat{y} = b_0 + b_1 x_1 + \ldots + b_p x_p$

Bernoulli regression $\quad \phi^{-1}(\hat{p}) = b_0 + b_1 x_1 + \ldots + b_p x_p \quad$ e.g. $\phi^{-1}(\hat{p}) = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ for logistic link

Poisson regression and Neg Binom regression $\quad \hat{y} = e^{b_0} e^{b_1 x_1} \cdots e^{b_p x_p} \implies \ln(\hat{y}) = b_0 + b_1 x_1 + \ldots + b_p x_p$

Weibull regression $\quad \hat{y} = \Gamma\left(1 + \frac{1}{\hat{k}_{MLE}}\right) e^{b_0} e^{b_1 x_1} \cdots e^{b_p x_p} \implies \ln(\hat{y}) = \ln\Gamma\left(1 + \frac{1}{\hat{k}}\right) + b_0 + b_1 x_1 + \ldots + b_p x_p$

Cox PH regression $\quad h(t) = h_0(t) e^{b_0} e^{b_1 x_1} \cdots e^{b_p x_p} \implies \ln(h(t)) = \ln(h_0(t)) + b_0 + b_1 x_1 + \ldots + b_p x_p$

WLOG, consider $j=1$, ~~the first covariate, and let it be~~ "blood sugar". let it be measured in mg/dL.

Here's the causal interpretation for OLS for x_j = blood sugar measured in mg/dL and y = blood triglycerides measured in mmol/L for $b_j = 1.17$ and $s_{b_j} = 0.25$

"If blood sugar (name of metric x_j) is increased (manipulation!) by one mg/dL (unit of metric x_j) and all other measurements remain constant, blood triglycerides (name of linear response) will ~~accordingly~~ increase ("increase" if $b_j > 0$ o/w "decrease") by an estimated $1.17 \pm .25$ (value of $b_j \pm s_{b_j}$) mmol/L (unit of response y) assuming the blood triglyceride level is linear in the covariates (the response function assumption)

Focus is on the manipulation / intervention of x_j

Let's do this for logistic regression where x_j = blood sugar (mg/dL) and $y = 1$ if person will get diabetes. and $b_j = 0.49$, $s_{b_j} = 0.12$

"If blood sugar is increased by one mg/dL and all other predictors remain constant, the log odds of getting diabetes will roughly increase by an amount 0.49 ± 0.12 assuming the log odds of getting diabetes is linear in these p covariates."

How about for Weibull? x_j = blood sugar (mg/dL) and y = mo. survival (yr) and $b_j = -0.02$, $s_{b_j} = 0.007$

"If blood sugar is increased by one mg/dL and all other predictors remain constant, the log survival will roughly decrease by an amount 0.02 ± 0.007 log(yr) assuming the survival is Weibull distributed with log mean linear in these p covariates."

Alternately, if we want to use survival in years, we can use the proportional multiplicative interpretation. $e^{b_j} = e^{-.02} = 0.98 \Rightarrow -2.0\%$. For s_{b_j}, more difficult, read delta method

"If blood sugar is increased by one mg/dL and all other predictors remain constant, the survival will roughly decrease by an amount 2.0%. assuming ..."

prop. hazard assumption

For coxPH, same covariates

"If blood sugar is increased by one mg/dL and all other predictors remain constant, the log hazard rate of survival (at any time) will roughly decrease by an amount 0.02 ± 0.007 assuming the survival process has a hazard rate log linear in the p covariates.